DOCUMENT RESUME

ED 403 888 IR 056 260

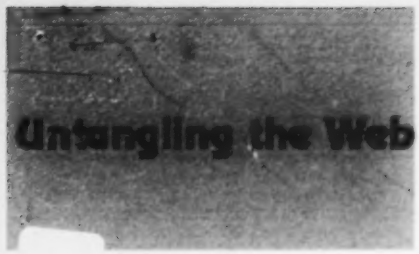
AUTHOR Lager, Mark A.
TITLE Spinning a Web Search.
PUB DATE 96
NOTE 10p.; A downloaded version of one of 29 conference papers (for those in ERIC, see IR 056 258-274) placed on the Internet under the following overall title: "Untangling the Web. Proceedings of the Conference Sponsored by the Librarians Association of the University of California, Santa Barbara, and Friends of the UCSB Library" (Santa Barbara, CA, April 26, 1996).
AVAILABLE FROM Electronic version: http://www.library.ucsb.edu/untangle/
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Artificial Intelligence; Computer Networks; Directories; *Indexing; *Information Retrieval; Information Seeking; Information Services; Internet; Natural Language Processing; *Online Searching; Reference Services; *Relevance (Information Retrieval); *World Wide Web
IDENTIFIERS Query Processing; *Search Engines

ABSTRACT

This paper focuses on techniques for retrieval used in information sciences and in World Wide Web search engines. The purpose of reference service and information science is to provide useful information in response to a query. The two metrics of recall and precision serve to express information retrieval performance. There are two major categories of searching tools on the Web: directories and search engines, and both require an indexing system. Building an index is done by either human or computer; for computers, the software program called a robot, or spider, or wanderer, visits each site and gathers information. Search engines on the Web incorporate a number of techniques to assist in both recall and precision. New trends in information retrieval include: artificial intelligence, probabilistic logic, query-by-example, query expansion, automatic summaries, natural language processing, which is a part of artificial intelligence, and concept-based searching. A description of the search engines available from Netscape's Net Search is provided at the end of this paper. (Contains 26 references.) (AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Mark A. Lager
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 403 888

Spinning a Web Search

Mark Lager
California Lutheran University

Copyright 1996, Mark Lager. Used with permission.

Abstract

Thesauri, subject headings, and keyword searching have been the customary strategies used for location of materials. Categorizing by subject or descriptor term creates a thesaurus of specified terms, a controlled vocabulary.

Searching this list identifies cataloged items. However, new computing techniques- artificial intelligence, natural language processing, relevancy feedback, query-by-example, and concept-based searching- have added a new sophistication to information retrieval. Robots, computers that search the web using these new techniques (AKABoolean searches spiders or wanderers), offer the 'net searcher a higher degree of precision in retrieval.

The presentation is targeted toward WEB searchers, in particular, reference librarians and those who navigate the Internet on a frequent basis. This presentation will look at search engines, comparing search techniques and noting differences. The workshop will identify use of new computing strategies for information retrieval within each engine.

"Spinning a Web Search" requires an electronic classroom with large screen projection for Netscape (or other browser), or classroom demonstration will focus the workshop: handouts (with exercises) will be provided for clarification and training. The workshop will be one hour.

Introduction

Information retrieval has always been the focus of information scientists and reference librarians, trying to provide relevant materials in response to a user's query. As reference librarians, we often pride ourselves on knowing our collections and what it holds. Or, we sleuth it out, knowing where to find the information. We use major indexing and abstracting tools, or our library catalogs, or online services to discover the needed information. It has been a domain left to those who can ferret out, or research, to uncover the user's information need. The World Wide Web (WWW) has burst that bubble! No longer is it a domain reserved in the library, or a special collection. The Internet, and especially the WWW, has made a major collection of information accessible to everyone, anywhere. The Web uses hypertext, a protocol or common language, to jump easily between files; the WWW opens the publishing arena to anyone with a computer. These hypertext links join databases, files, sounds and pictures; texts, library catalogs,

IR056260

BEST COPY AVAILABLE

songs, video, and more are now available to the computer-literate. Unlike the orderly world of the library collection, this new source of information, is chaotic, often not organized and includes information not of high quality. As Brian Pinkerton states, "The World Wide Web is decentralized, dynamic and diverse; nativagion is difficult and finding information can be a challenge." (Pinkerton, 1994). The useful and the innocuous are lumped together in this huge collection. Academic information (e.g., journal articles and course materials) is combined with social culture information and with personal home pages. There is no separation. Mark Nelson calls this information anxiety - the overwhelming feeling one gets from having too much information or being unable to find or interpret data. (Nelson, 1994). To be of any information value, the data must first be organized and retrievable, providing some structure. Search tools have begun to put some organization to these uncharted waters. Current trends in information retrieval offer better opportunities to make more efficient use of this information resource.

This workshop is designed to explain trends in retrieving information. The workshop will focus on techniques for retrieval used in information sciences and in WWW search engines. We will look at some search engines and their make-up to view techniques used. It is not designed to explore search strategies of the various search engines (i.e., + for plurals, - for negations, - for phrases, etc.).

Recall vs. Precision

The purpose of reference service and information science is to provide useful information in response to a query. Despite the methods used, whether I use my knowledge to know where to locate the information, or whether the computer searches its index, the information retrieved can be classified in one of two categories of measurable statistics: whether the information retrieved is considered relevant, or whether all the relevant material was retrieved. These two metrics of recall and precision serve to express information retrieval performance.

Recall is percentage of total relevant documents retrieved from all documents. Recall refers to how much information is retrieved by the search. Total recall would locate every document that matched the search criteria in a database. Precision is the percentage of documents retrieved that the searcher is actually interested in. Precision focuses on the relevant, most useful items retrieved in the search. Recall with high precision is the ultimate goal. The goal of information retrieval scientists is to provide the most precise or relevant documents in the midst of the recalled search results. (Dataflight, 1995)

Let me use three examples to illustrate recall, and then precision. When GM does a recall of its cars, it takes back out of the total number of GM manufactured cars, those that are of a certain model or year. I am searching for all red books in my collection of 1000 books. I pull from the shelves 50 books. Recall, is 50 out of 1000 or 5%. This figure recalls from the collection only those items that match my need. %. If I am looking for disk brakes, I may begin to search the index with the word "brakes." I get a high recall that matches my query since it includes disk drives and disk brakes.

Precision focuses on only the relevant documents out of the recalled items. For the GM example, precision is the number of recalled cars that actually have the defective part. This means they may have to check all the recalled cars to only find a few with the defect. Out of the 50 red books I collected, there are 25 that match the shade of red that I want. For the disk drives, who knew what kind of disk I meant? I must reinitiate the search to qualify my need.

The key phrase is "relevant." This is the quandary in information science and for librarians. Who determines what is relevant? I know what shade of red I want, but who else knows that? I know what kind of brakes I want, so why doesn't the computer read my mind? GM knows which cars have the problem, why don't they just let the owners crawl under the car to see if the part is defective?

Relevance

Traditional finding aids to assist in information retrieval have focused on Subject Headings, descriptors, or established vocabulary. (Examples include LCSH, ERIC thesaurus, UMI's Controlled Vocabulary List). Items cataloged using MARC format is one primary means to retrieve the item;. in fact, the MARC format is an information retrieval standard: Z39.2. (Bowker, 1991) The subject headings are used to determine the primary focus of a book/article, i.e. to help precision if the user makes use of the established subject heading. In the 70's, automated

catalogs were created to use the traditional access points - author, title, series, subject and added entries. Over the years, catalogs were enhanced by adding search capabilities to additional fields - e.g. 5xx (notes) fields. More data, higher recall. Current online catalogs offer keyword searching and Boolean searching to assist in precision. By adding more word access, it is hoped that the searcher's terms will be in the index. With more data to search, the search engine can return more documents gaining greater recall. It is for the user to sift through the recalled documents to find useful ones. By adding pointers between words, builidng interrelationships, an index can becomes highly beneficial for retrieval of information. (EB, 1995)

The Web

The World Wide Web began in 1989 at the CERN Particle Physics Lab in Switzerland. The Web, however, did not gain widespread popular use until browsers like NCSA Mosaic became available in 1993, and Netscape in 1994. The task of making the Web more searchable began soon thereafter with search tools as the Wanderer and JumpStation in 1993. The Web doubles every five months which makes indexing and updating a formidable task. (9) With over 3 million host computers on the Web, (Pike, 1995) it is difficult to imagine finding only relevant, precise information in the midst of all the data. The computers on the Internet share a common language called TCP/IP. Based on this commonality, computers communicate with each other. One method for communication that is used on the Internet is called client/server. Think of this as "one speaker at a time" or only one computer "has the floor." Of course, with the high speed, it looks like conversations are simultaneous. Just as a patron asks a librarian for assistance on an information request, so the client (piece of software) asks the librarian (server that houses the information) who responds. The information, however, is only as good as what is indexed, retrievable.

Web Indexing

There are two major categories of searching tools on the Web: directories (what we know as an index) and search engines. Both require an indexing system. Building an index is done by either human or computer. For computers, the software program, called a robot, or spider, or wanderer, visits each site and gathers information. A robot gains the home page address (URL) and then recursively visits some or all of the links. The index contains URLs, titles, headings, and other words from the HTML document. (HTML is an acronym for hyptertext mark-up language). Each index is different depending on what information is deemed important, e.g., Lycos only does top 100 words, first 20 lines of text. AltaRelevance - marked with red X Vista and Infoseek index every word. Robots perform, resource discovery, the term for a robot's ability to summarize and index the data on the Web, to automatically update information and change or eliminate dead links. (Fischer, 1995)

Robots work in either of two ways:

> Depth first-- these examine a Web page and follow each link in as much depth as possible then follow the links to other external Web pages. Lycos is the best example of this.
> Breath first-- these collectors start at the top level of a Web page and follow all of the top level links to other Web pages.

Acting like a sophisticated web browser, the robot automatically retrieves documents or other information until told to stop. (Boutell, 1995) Creating the index is usually done by a robot, since that is the more efficient means of searching. The robot, a simple executable program, sweeps a portion of the web, retrieves, parses, and stores pieces of the HTML document and then reindexes the data. There are over 30 robots in existence. (Fischer, 1995) Below is a listing of a few robots:

Explorer Katipo Titan Aretha Northstar

Python Htmlgobbler Pagemator Websmurf Lycos

Arachnophillia Scooter Webforager

Directories, like Yahoo, Infoseek Guide or Librarian-Built Subject Guides , or John Makulowich's Awesome List , are examples of a created list by subject. These can be created by machine or by human. The searcher can click on the

topic and see a listing of sites. Directories are an excellent place to start, and I recommend users to begin with these, especially if they are not familiar with the Web. Just as we would send persons to begin a search to the printed index volumes to look up the topic, so the online directory lets the user browse among a reviewed list. Indexes provide a clear condensed grouping by subject, saving time. However, lists are different since they have been created using specific criteria of the indexer. There is no standard for Directory terms.

Search engines

The search engine provides more control for the user in performing a search. Engines use the index to fetch terms of the query. This means that the more data in the index, the higher the recall. Indexing every word or the most used words can lead to higher recall depending on the search query. The larger the index, the more possibility of hitting upon the words of the query. And, with the size of the Web, the more often the index is updated, the greater the number of hits.

Search engines on the Web incorporate a number of techniques to assist in both recall and precision. There are search engines that employ traditional methods like thesauri or Boolean searching. Rather than being only a keyword search, the engine will make logical connections to a thesaurus to enhance recall. Using Boolean logic (and, or, not, adjacency operators) search engines can assist in making the query more precise. Different engines have different defaults.

Natural Language Processing: Relevancy feedback/weighing

probabilistic logic: query by example

fuzzy logic: query expansion

Bayesian networks: case-based reasoning

parallel computing (Inktomi): concept based searching

(For a listing of the variety of search engines, over 120, see
http://ugweb.cs.ualberta.ca/~mentor02/search/search-all.html)

New Trends in IR:

Artificial Intelligence

AI, the capacity of a digital computer or computer-controlled robot device to perform tasks commonly associated with the higher intellectual processes characteristic of humans, such as the ability to reason, discover meanings, generalize, or learn from past experience. The term is also frequently applied to that branch of computer science concerned with the development of systems endowed with such capabilities. (EB, 1996) Artificial intelligence refers to creating computers that can think and reason. AI focuses on finding a logical, mathematical way to represent knowledge. The computer can be programmed with this mathematical model to assist in decision making, information retrieval, and analysis. Then, when a query is asked, the computer follows the rules for a response. AI has many facets, including robotics, expert systems, and voice recognition and simulation. Search engines incorporate some of the fascinating trends in AI.

Probabilistic Logic

Will it rain today? What is the possibility of my car needing an oil change? Or, what is the chance of getting an A on my history test?. There are many questions like these that cannot be answered with an affirmative or negative answer. Uncertainty reigns. In an effort to make a decision which accounted for such doubt, in the midst of chaos, a branch of logic was defined to study probability. Since the 16th and 17th centuries, **probability theory** has been used to explain chance. Such questions rely on a factual information as history coupled with probability. In information

retrieval, the same applies. By setting up a formula, an algorithm, that places values on words, their interrelationships, proximity, and their frequency, the computer can be used to help locate relevant sites. By computing these terms together, the search engine can produce a relevancy ranking that is then displayed to the user. (De Bra, 1995)

Probabilistic logic is founded on the presumption that certain factors can be established logically and mathematically to focus a search. It is similar to fuzzy logic where the central notion is that truth values (in fuzzy logic) or membership values (in fuzzy sets) are indicated by a value on the range [0.0, 1.0], with 0.0 representing absolute Falseness and 1.0 representing absolute Truth. (Brule, 1985)

One method of explaining possibilities was created by Rev. Thomas Bayes, a mathematician from the 18th century. His theorem tried to apply a mathematical, logical representation to various factors. Here is an example of his mathematical model of probability (Case, 1995):

p(hle,i)=p(hli)*p(elh,i)/p(eli)
p=probability
h=hypothesis
e=evidence
i=context

Returning the value of the possibility is called weighting. Weighting of terms is based on a number of factors, as used by the search engine: A. *Relative frequency*, the more times a word/phrase appears the more weight it carries. The frequency of the term places a higher weight on he document. B. *Closer to top* - documents that have the query term(s) in the URL or in the title are weight more strongly. Terms appearing in the top of the document is weighted more relevant than at the bottom. C. *More occurrences* - if a document uses the key terms often, it is ranked more highly than one that seldom uses that particular term. D. *Adjacency or proximity* - words from the query that are found next to each other in the document score higher.

Query by example

Query-by-example (QBE) is the concept of providing the search engine an example for which to Using this example, the system returns other like documents. For example, I want a book about gorillas, published in 1984, that has a green cover. I have set up an example of what I am looking for using all my qualifications. Search engines use the technique to set up queries to find similar pages or files. The search is reinitiated using the example as the new source for the query. This interactive searching gives the user more control over the search process. Users can find more documents like the one selected. The results returned are then more focused because of the qualified terms. (Sugihara, 1995)

Query Expansion

Once a search has been completed, it often tends to need to be enhanced or changed. A library patron who comes to the desk asks one question, but usually there is some other additional information need. The purpose of the librarian is to elicit that actua l request. The quest of the information scientist is to discover how the computer can assist in evoking that query and its modifications. Newer search engines provide the user with more control over the query, by adding a means to resubmit the search with any changes.

Automatic Summaries

Many search engines incorporate a feature that creates summaries of the document retrieved. This can be based on taking information from the first few lines, or by locating key statements from within the document.

Natural Language Processing

Natural Language Processing is the act and science of getting computers to understand natural language. It is a part of artificial intelligence. (Case.) Computers process language not only by exact match, using keywords. NLP

involves using a set of concep ts to sort out the interrelationships of words. The computer breaks apart the sentence into its semantic parts: nouns, verbs, adjectives, etc., and then it creates links. Since language can be ambiguous, vague, or metaphorical. NLP seeks to compute the relationships between words, giving each a correlate to the words around it. Put into a formula, the computer then makes assumptions based on its logic. Although similar to a keyword search, the search engine allows a user to make the query as if asking a librarian.

Concept-based searching

Using the idea of a thesaurus, a search engine can expand upon the keyword that a user may input. In this manner, users do not have to know the exact words to use to retrieve relevant documents. And, instead of reinstituting the search based on "confidence" or "weighting," the search engine automatically includes the like terms.

Search Engines

A survey of the Search Engines available from Netscape's Net Search will help in explaining some of the techniques discussed. By conducting a search for current trends in information retrieval, differences can be seen in the structure and techniques of each engine.

Alta Vista http://altavista.digital.com/

Techniques and features
Boolean - must use and, or, not, near (10 words) in Advanced Search
Allows user-influenced results ranking
Ranking: title words or first few words

Closer to each other
Document has more of the words
More copies of the words throughout

Parentheses for nesting
Can restrict to field (qualifiers)

Excite http://www.excite.com/

Techniques and features
Concept based searching-use statistical strength of interrelationships between words
Creates its own knowledge base (or internal thesaurus)
QBE - "similar documents"
Boolean searches
Keyword searches
Relevance - marked with red X
Robot is called Architext

Infoseek http://www.infoseek.com/

Techniques and features
Weight terms (required, desirable, undesirable)
Similar pages - QBE
Boolean operators
Natural language
Search mechanisms

BEST COPY AVAILABLE

Lycos http://www-msn.lycos.com/

Techniques and features

Probabilistic retrieval
Indexes top 100 words and 20 lines of abstracts
Keyword searching
Boolean searching
Automatic truncation
Adjacency 0.0 - 1.0
Results categorized
Terms in bold
Relevancy: early on vs. farther down

Magellan http://www.mckinley.com/

Techniques and features
Reviewed by writers
Boolean searching
Green light for information for all age groups
Web, ftp, gopher, newsgroups, telnet sites
Browse directory or Use search engine
Relevancy = frequency of words
Browse button
Robot named Verity
Lists up to 20 pages at the bottom of the screen

Open Text http://www.opentext.com/omw/f-omw.html

Techniques and features
Boolean searching
Field operators: anywhere, summary, title, first heading, URL
Query-by-example

Conclusion

Information search and retrieval is of major importance in locating relevant materials. The ability to aid and assist a user in finding relevant information is the goal of librarians and information scientists. On the Web, search engines have made the pr ocess easier by incorporating a number of newer techniques which include artificial intelligence, Bayesian statistics and probability theory, weighting, and query by example. With the goal of finding relevant materials, these new techniques locate infor mation and also refine the search query. Since search engines have different criteria in creating the indexes, it is most useful to use more than one engine in searching the Web to gain relevant information. As a rule, the more critical or focused the q uery, the more engines should be applied. With advances in the tools for information retrieval, the future holds exciting possibilities for searching on the World Wide Web.

Bibliography

"Alta Vista: Tips". [http://altavista.digital.com/cgi-bin/query?pg=tips] 1995.

Birnham, L. "Natural Language Processing". [http://yoda.cis.temple.edu:8080/nlp/nlp-course/lecture1]. 1994.

The Bowker Annual: Library and Book Trade Almanac. 35th edition. 1990-1991. New York: Reed Publishing, c. 1990.

Boutell, Thomas. "World Wide Web FAQ. Robots." [http://www.glocom.ac.jp/mirror/sunsite.unc.edu/boutell/faq/robots.html] 1995.

Broule. Thomas. "World Wide Web. FAQ."
[http://www.glocom.ac.jp/mirror/sunsite.unc.edu/boutell/faq/robots.htm]. 1995.

Brule, James F. "Fuzzy Systems- A Tutorial." [http://www.csu.edu.au/complex_systems/fuzzy.html]. 1985.

Case, J. "Natural Language Processing". [http://bones.wcupa.edu/~jcase/ciir1a-report.html]. 1985.

Dataflight Software, Inc. "Concordance Information Retrieval System."
[http://concord.lax.primenet.com/Whitepap.htm]. 19 February 1996.

De Bra, Dr. P.M.E. "Hypermedia structures and systems."
[http://www.win.tue.nl/win/cs/is/debra/cursus/retrieval.html]. 1996.

Excite. "Handbook: NetSearch."
[http://www.excite.com/cgi/comsubhelp.cgi?display=html;path=/query.html;section=search;Help=Help]. 1996.

Encyclopedia Britannica. [http://www.eb.com:180/] 1996.

Fischer, Keith. "Preliminary robot.faq. [http://info.webcrawler.com/mak/projects/robots/active.html] 6 Nov. 1995.

Gray, Matthew. "Measuring the Growth of the Web". [http://bokonon.mit.edu/growth/growth-report.html]. 1995.

"Infoseek Tips." [http://guide.infoseek.com/] c.1996.

Koch, Traugot. "Robot-based WWW Catalogs." [http://www.ub2.lu.se/nav_menu.html#robo]. 1996.

"Magellan Frequently Asked Questions." [http://magellan.mckinley.com:80/mckinley-txt/250.html#howperform] 1995.

Needleman, Mark. Information Retrieval and the ASISKeyword searches Standards Committee. *American Society for Information Science Bulletin.* Feb 1995, 21(3), p. 25-26.

Nelson, Mark R. "We Have the Information You Want, But Getting It Will Cost You: Being Held Hostage by Information Overload." [http://www.acm.org/crossroads/xrds1-1/mnelson.html]. Sept 1994.

Notess, Greg R. "Searching the World Wide Web: Lycos, Webcrawler, and more." *Online,* July 1995. 19(4), p.48-53.

Pike, Mary. *Using the Internet.* Second edition. Indianapolis, IN: Que, 1995

Pinkerton, Brian. "Finding What People Want: Experiences with Webcrawler."
[http://info.webcrawler.com/bp/WWW94.html]. 1994.

Sugihara, J. ICS421 Lecture Notes 1. [http://www.ics.hawaii.edu/~sugihara/course/ics421s95/note/3-06n13]. Jan. 1995.

Van Rijsbergen, C. J. "Information Retrieval." [http://www.dcs.glasgow.ac.uk/Keith/Preface.html].

"WebCrawler Help." [http://wc2.webcrawler.com/WebCrawler/Help/Examples.html]. 1996.

Winship, Ian R. "World Wide Web searching tools, an evaluation."
[http://www.bubl.bath.ac.uk/BUBL/IWinship.html]. 1995

"Yahoo! Help." [http://www.yahoo.com/docs/info/help.html]. 1996.

Direct questions or comments about this page to webmaster@library.ucsb.edu .

Search "Untangling" Papers | "Untangling the Web" Home | InfoSurf Home

This is an official University of California, Santa Barbara Library web page.